A comparison of statistical models for short categorical or ordinal time series with applications in ecology

Noëlle Bru, Laurence Despres, Christian Paroissin

A comparison of statistical models for short categorical or ordinal time series with applications in ecology

Noëlle Bru [*] Laurence Despres [†] Christian Paroissin [‡]

Last version: 23rd February 2007

Running title: Statistical models for categorical data time series in ecology.

Abstract

We study two statistical models for short-length categorical (or ordinal) time series. The first one is a regression model based on generalized linear model. The second one is a parametrized Markovian model, particularizing the discrete autoregressive model to the case of categorical data. These models are used to analyze two data-sets: annual larch cone production and weekly planktonic abundance.

Keywords: time series, categorical variable, ordinal variable, regression model, Markov chain, auto-regressive process, estimation.

AMS 2000 Classification: 62M10, 62M02, 62M05

[*]Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques Appliquées - UMR CNRS 5142 / IUT STID, Avenue de l'Université, 64013 Pau cedex, France. Email: nbru@univ-pau.fr

[†]Université Joseph Fourier, Laboratoire d'Ecologie Alpine (LECA) - UMR CNRS 5553, BP 53, 38041 Grenoble Cedex 09, France. Email: laurence.despres@ujf-grenoble.fr

[‡]Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques Appliquées - UMR CNRS 5142, Avenue de l'Université, 64013 Pau cedex, France. Email: cparoiss@univ-pau.fr

1 Introduction

The quest to understand mechanisms behind the temporal dynamics of a natural population (animal or plant) always yields useful information for ecological biodiversity management. The present work was motivated by the analysis of time-fluctuations of an ecological time series, the annual larch cone production. Because of the impracticability of quantitative evaluation of such population size, only semi-quantitative data are often available numerically. Data are coded with finite ordered categories or levels. From this some natural questions arise. Among them, the following one is of crucial interest here: do lagged values determine future production? The same basic problems occur as for classical quantitative time series but here the greatest difficulty stems from the nature of the studied process (as previously mentioned, working on categorical variables induces many difficulties since most of the notions used for quantitative variables have no more sense in such context).

Statistical models have been useful instruments for testing hypothesis concerning the mechanisms behind temporal evolution and to characterize temporal patterns. Two models are used throughout this paper to achieve this goal: a regression model (Fokianos and Kedem, 2002, 2003) and a parametrized Markovian model (Jacobs and Lewis, 1978a). The first one is a regression model for categorical time series which is based on generalized linear regression theory (McCullagh and Nelder, 1989). Such model extend linear models to accommodate both non-normal response distributions (which is the case in the study of categorical data) and transformations to linearity. So, applying a generalized linear models consists in two choices: a family of probability distribution and a link function between the response and the predictors. For categorical data some widely used models are: multinomial-logit (Agresti, 1990) and cumulative odds models (McCullagh, 1980). Adaptation of such models to categorical times series is easy to do putting past observations at different lags as categorical predictors of the response at time t (Fokianos and Kedem, 2002, 2003). The second one is indeed an adaption of the discrete auto-regressive (DAR) model introduced by Jacobs and Lewis (1978a) to the case of categorical time series. As noticed by McKenzie (2003), DAR models ¡¡ would be more suited to modelling dependent sequences of categorical observations, but this does not seem to have been attempted yet ¿¿. To the best of our knowledge, no advance in this direction is made since the paper of McKenzie.
These two models have some advantages and some disadvantages which are not necessary the same, implying a complementarity between these two approaches. Among the common advantages, the main one is that they are easy to be interpreted by the practitioners. Since most of the time series in ecology are short-length (for a statistical purpose), we have to consider only models involving a reasonable number of parameters. That is the reason why we will focus on one order lagged model (even these models can be extended easily to large order lag values). Among the inconvenient of the DAR model, the main one is the stationarity of the time series, which can not be checked by any statistical tests (see (McGee and Harris, 2005) for a discussion about several notions of stationarity for categorical time series). However it allows us to derive a simple model for taking into account missing values (a contrario to the regression model, the DAR can not treat directly the case of missing values). Our approach differs highly of the one recently proposed by Bandt (2005). Indeed he considers a continuous-state, but non-Gaussian, time series and its analysis relies only on the ordinal property of $\mathbb{R}$. Moreover his methodology requires a long time series, which is not realistic in many real cases.

Motivation of the present work is the analysis of time series of larch cone production data in spatially disjoint locations in order to determine some temporal patterns of larch cone production dynamic at different locations and to discuss some kind of spatial synchrony. Data are detailed in the first section. Next section 3 is devoted to present two regression models: one for categorical time series and one for ordinal time series. These two regression models have been studied by Fokianos and Kedem (2002; 2003). In section 4 we adapt the one order discrete auto-regressive model to the context of categorical data (the ordinal characteristic is not taken into account in this model). In particular we develop independence tests and estimators of the various parameters of the model. In section 5, we apply these models to two real data sets: the first one deals with annual larch cone production (over 31 years) and the second one with weekly planktonic abundance (during one year). Last section is devoted to conclusion and discussion.

2 Motivations

The masting is the intermittent synchronous production of seed crops by a plant population (Kelly and Sork, 2002). It often shows an evolved strategy related to others environmental masting patterns such as rainfall, temperatures, ... Thus variability in seed production according to past values is a good descriptor of environmental changes in climate for example. The information arising from the characterization of temporal patterns on such time series could be used to infer role of environmental parameters and other mechanisms (Price *et al.*, 2006). The data accounting cone production were registered for 31 years on four valleys located in the Southern French Alps (in the same area of the Alps called "Briançonnais"). Here we will consider four different sites selected to be at comparable altitudes (ranging from 1800m to 2200m): Ayes (altitude: 2200 meters), Montgenèvre (altitude: 2200 meters), Névache (altitude: 1800 meters) and Prorel (altitude: 1800 meters). Cone production at a given site was roughly estimated at the beginning of the cone development by counting cones along one meter of branch for at least one hundred randomly selected trees. The intensity of larch cone production at any site was then classified into six classes (Roques, 1988) from no cones (coded 0) to very heavy crop i.e. more than two hundred cones per tree (coded 4). Annual cone production is considered to be the realization of an ordinal time series with values $\{0, 0.5, 1, 2, 3, 4\}$ corresponding to a scale classification endowed with a natural ordering. Data are plotted on figure 1.

When studying the dynamic of the larch population on each sampling sites, a first step could be to identify temporal patterns of cone production and then to compare each patterns from one site to others to conclude or not at a spatial synchrony on a "short" regional spatial scale (Liebhold *et al.*, 2004). However, the observed series in figure 1 do not exhibit obviously the presence of such patterns. The salient features of the series are: no seasonality, high location to location variability with respect of duration and beginning of intensive larch cone production, presence of missing values, ... However visual remarks should be considered carefully.

Such time series could appear as too short-length for the statistician who generally needs a lot of information to infer on a phenomenon but the data are collected from 1975 to 2005, which corresponds to an entire career of a biologist!

3 Regression models for categorical and ordinal time series

The model used here is a generalization of classical regression models to the case of time-dependent categorical observations and was studied by (Kauffmann, 1987) (see also (Fokianos and Kedem, 2002, 2003) for a good summary of the main theoretical aspects).

3.1 Introduction to generalized linear models for qualitative time series

Assume that the observed series is a particular realization of the stochastic process in discrete time $\{Y_t\}$ which will be described below. Values of Y_t are supposed to belong to a finite set $E = \{1, \ldots, k\}$ of k ordered or not categories. Because we are interested in temporal dependence between successive observations, we condition on the observed past. For any positive integer l, let us denote by $\mathcal{F}_{t-l}$ the σ-field generated by $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-l}$. Let $\mathbf{Y}_t = (Y_{t,1}, \ldots, Y_{t,k-1})'$ where $Y_{t,j}$ equals to 1 if the j-th is observed at time t and 0 otherwise. The analysis of time series based on a generalized linear model suppose that the response variable is influenced by its past values which are viewed as predictors influencing the distribution of Y_t by way of a transformation of a linear combination.

$$\mathbb{E}\left[Y_t|\mathcal{F}_{t-l}\right] = \mathbf{h}\left(\mathbf{Y}'_{t-l}\beta\right),$$

where l is the order of the lag time and $\mathbf{Y}'_{\mathbf{t-l}}$ is the covariate matrix containing the lagged values of the response variable until lag l. In the following we will focus on $l \in \{0, 1, 2\}$ (since we aim at treating short-length time series). The vector β is a vector of time-invariant parameters to be estimate which will reflect the intensity of the

dependency between the response and its past. Because the response variable is a categorical time series, we have the following relation:

$$\pi_{t,j,l} = \mathbb{E}[Y_{t,j}|\mathcal{F}_{t-l}] = \mathbb{P}(Y_{t,j} = 1|\mathcal{F}_{t-l}),$$

for every $j \in \{1, \ldots, k\}$ and every $t\{1, \ldots, n\}$, where $\pi_{t,j,l}$ is a transition probability. Let $\pi_{t,\cdot,l} = (\pi_{t,1,l}, \ldots, \pi_{t,k-1,l})$. Some adequate regression models for categorical data falls in the family of generalized linear models which links vector of transition probabilities of the response vector to the covariate process through the equation:

$$\pi_{t,\cdot,l} := \pi_{t,\cdot,l}(\beta) = \mathbf{h}(\mathbf{Y}'_{t-l}\beta) \qquad \text{or} \qquad \mathbf{h}^{-1}(\pi_{t,\cdot,l}(\beta)) = \mathbf{Y}'_{t-l}\beta \; . \tag{1}$$

In other words, the study of having the response $Y_t = j$ at time t is equivalent to carry out a regression on covariates which are the lagged values of the categorical response process. This model is also called a Markov regression model for categorical time series. The function $\mathbf{h}$ is called the inverse link function and is related to a link function that describes how the mean depends on the linear predictors. For each response distribution there exists a variety of link functions to connect the mean with the linear predictor. The use of a generalized linear model is the choice of a combination of response distribution and a link function.

3.2 On the choice of the link function

The link function should adapted to the type of data (Fokianos and Kedem, 2003):

- Nominal data: the most commonly used model for categorical (or nominal) data is the multinomial logit model (Agresti, 1990):

$$\pi_{t,j,l}(\beta) = \frac{\exp(\beta'_j y_{t-l})}{1 + \sum_{q=1}^{k-1} \exp(\beta'_q y_{t-l})} \; , \tag{2}$$

  for any $j \in \{1, \ldots, k-1\}$. This equation also defines log-odds ratios relative to π_{tm} by:

$$\log\left(\frac{\pi_{t,j,l}}{\pi_{t,k,l}}\right) = \beta'_j \mathbf{y}_{t-l} \; . \tag{3}$$

- Ordinal data: since data are ordinal, its is more convenient to model the cumulative probability function of Y_t. For ordered categorical time series a reasonable choice of link function is the logistic distribution one which leads to the proportional odds model (McCullagh, 1980):

$$\mathbf{h}^{-1}(x) = \frac{1}{1 + \exp(-x)} \; . \tag{4}$$

  It follows that the link function is:

$$\log\left(\frac{P\left[Y_t \leq j|\mathcal{F}_{t-l}\right]}{P\left[Y_t > j|\mathcal{F}_{t-l}\right]}\right) = \mathbf{Y}'_{t-l}\beta \; . \tag{5}$$

3.3 Parameters estimation and global adequacy criteria

Since the joint distribution of response and covariates is often not easy to establish, the likelihood methods are not applicable to estimate the vector of regression coefficients β. As we are interested in the estimation of the effects of the covariates on the response, we can use the inference theory based on partial likelihood function. The reader can refer to (Fokianos and Kedem, 2002; Viennet *et al.*, 1998) for more details and application. The partial likelihood method leads to non linear equations system. Multinomial models were fitted using the function `multinom` from library section `nnet` on `R`. Proportional-odds logistic regression models were fitted using the function `polr` from library section `MASS` on `R`. The vector of parameters of this model β is estimated using an iterative weighted least squares `IWLS` (Chalmers and Hastie, 1992; Venables and Ripley, 2002).

The analysis of the global adequacy and goodness of fit of such models to the data is discussed using the Akaike's information criterion (AIC) which also allows to compare several models. The values of this criterion depends on

the number of model parameters and penalizes models with large number of parameters. Such consideration is important in the study of short time series where the number of parameters can be rapidly equal to the length of the time series. The chosen model is the one which minimizes the value of AIC among the others.

In this preliminary work, no detailed analysis of the residuals of the models is done. Such analysis is important to assess the goodness of fit between the chosen model and the observed data but was not the priority of this paper.

4 Discrete auto-regressive model and categorical data

The discrete auto-regressive (DAR) model introduced by Jacobs and Lewis (1978a; 1978b; 1978c) is used here to model categorical data. Some independence tests are developed, using either the Markov property or runs properties. Estimators of the parameters are studied in the precise context of categorical data. Simulated data are used to illustrate numerically the quality of these estimators.

4.1 Introduction and model

In a series of papers, Jacobs and Lewis (1978a; 1978b; 1978c) introduced and studied time series models for discrete variables. Among them we will focus here on the discrete auto-regressive of order 1, denoted by DAR(1). Such process $\{X_t\}$ is a discrete-time stochastic process with values on a finite ordered set $E = \{1, \ldots, k\}$ and is defined as follows:

$$\forall t > 0 \ , \quad X_t = V_t X_{t-1} + (1 - V_t) Z_t \ ,$$

where $\{V_t\}$ is a sequence of iid Bernoulli random variables with parameter $\alpha \in [0; 1]$ and $\{Z_t\}$ is a sequence of iid random variables having the distribution π on E, the two sequences being independent. Moreover we will assume that X_0 is distributed according to the distribution π, implying that the process $\{X_t\}$ is stationary. The case of $\alpha = 1$ is not interesting since $X_t = X_0$, with probability 1, for any t. The case of $\alpha = 0$ means that the process $\{X_t\}$ is simply a sequence of iid random variables having distribution π. Hence the parameter α could be interpreted as follows: the nearest to 0 α is, the more ¡¡ independent ¿¿ the sequence $\{X_t\}$ is. Indeed, for all $h \in \mathbb{N}$, $\rho(h) = \alpha^h$ is the auto-correlation function of a DAR(1) process. Hence DAR(1) models can be used to describe a situation of short range dependency with high correlation. It is easy to prove that stochastic process $\{X_t\}$ as defined above is a Markov chains on E with transition matrix P given by the following equation:

$$P = \alpha I + (1 - \alpha) Q \ ,$$

where ${}^tQ = [{}^t\pi | \cdots | {}^t\pi]$. Such Markov chain admits obviously a unique stationary probability distribution which is π. One can easily deduce the h-th power of Q and P: for all $h \geqslant 1$, $Q^h = Q$ and $P^h = \alpha^h I + (1 - \alpha^h) Q$, illustrating one more times the role of α.

This stochastic process could be generalized to higher order leading to the DAR(p) model. In fact, these models appear themselves to be a special case of mixture transition distribution (MTD) model introduced by Raftery (1985). Thus DAR(p) can be viewed as an alternative to MTD model. According to Raftery, a MTD model fits better data in general than a DAR(p) one, especially for $p \geqslant 3$. However here we will prefer to use a DAR(1) model since it has the following advantages over the MTD model: 1) the two parameters α and π play different roles: α is related to the correlation whereas π is the stationary distribution; 2) these models involve generally a reasonable number of parameters (more parsimonious) especially when few data are available; 3) parameters could be easily interpreted by a practitioner. But the special case of DAR(1) model presents the disadvantage of being restrictive over the transition matrix.

Here we are interested on the use of such stochastic processes for modeling categorical variables (here the k different modalities are encoded by using the k first positive integers). It implies that many characteristics of these processes have no sense in such context, as it is the case for the auto-correlation function (see above). Thus estimators developed by Jacobs and Lewis (1983, see pages 28–30) cannot be used. Hence we address the

statistical problem of estimating the parameters in a DAR(1) model in presence of categorical data. Assume we observe $X_0, \ldots, X_n$ for a fixed value $n > 0$. First we will test whether $\{X_t\}$ is a sequence of independent random variables ($\alpha = 0$) or not. In a second step we will estimate all the parameters of the model: α and π. Then we propose a very simple model in order to consider the case of missing observations.

4.2 Independence tests

In this section we aim at testing whether $\{X_t\}$ is a sequence of independent random variables ($\alpha = 0$) or not. Two ways will be investigated. The first one will use the Markov property of the DAR(1) model and the second one will be based on runs property. Anyway, all along this section, the null hypothesis H_0 will be ¡¡ $\alpha = 0$ ¿¿ and the alternative hypothesis will be ¡¡ $\alpha \neq 0$ ¿¿.

χ^2 test based on the Markov property The following test is a classical test for Markov chain (see (Reinert *et al.*, 2000) for an illustration in DNA analysis context). We only use the fact that $\{X_t\}$ is a Markov chain, but not the particular structure of its transition matrix. Classical results on Markov chain inference leads to the following estimate for the transition matrix P:

$$\widehat{P}_{j,j'} = \frac{N^n_{j,j'}}{N^n_{j,\cdot}} \ ,$$

where $N^n_{j,j'} = \sum_{i=1}^n \mathbb{1}_{\{X_{i-1}=j, X_i=j'\}}$ and $N^n_{j,\cdot} = \sum_{j' \in E} N^n_{j,j'} = \sum_{i=1}^n \mathbb{1}_{\{X_{i-1}=j\}}$. In other words $N^n_{j,j'}$ is the number of jumps from state j to state j' and $N^n_{j,\cdot}$ is the number of visits of state j, in the sequence of observations $X_0, \ldots, X_n$.

The null hypothesis can rephrased as follows: $P_{j,j'} = P_{j,\cdot} P_{\cdot,j'}$, for any $(j, j') \in E^2$. Under H_0, the maximum-likelihood estimate of $P_{j,j'}$ is:

$$\widehat{P}_{j,j'} = \widehat{P}_{j,\cdot} \widehat{P}_{\cdot,j'} = \frac{N^n_{j,\cdot}}{n-1} \frac{N^n_{\cdot,j'}}{n-1} \ ,$$

where $N^n_{\cdot,j'} = \sum_{j \in E} N^n_{j,j'} = \sum_{i=1}^n \mathbb{1}_{\{X_i=j'\}}$. Hence one has to consider the following statistics C^2:

$$C^2 = \sum_{j \in E} \sum_{j' \in E} \frac{[N^n_{j,j'} - N^n_{j,\cdot} N^n_{\cdot,j'}/(n-1)]^2}{N^n_{j,\cdot} N^n_{\cdot,j'}/(n-1)} \ .$$

Theorem 4.1 *Under the null hypothesis,* $C^2 \xrightarrow[n \to \infty]{d} \chi^2_{(k-1)^2}$.

Some well-known practical restrictions exist in order to be able to apply this test. As example, one can require that $\widehat{P}_{j,j'} > 5\%$, for any $(j, j') \in E^2$.

Tests based on runs property Unfortunately we cannot compute the power of the previous test, that is the reason we will now consider a second family of tests. These tests will be based on runs property of the model. Runs in sequence of iid Bernoulli distributions are studied for a very long time: this problem seems to be considered for the first time by Abraham de Moivre in 1756 (problem LXXIV in his book *The Doctrine of Chances*). For an historical perspective, see the introduction of the Part I of (Mood, 1940). Most of the existing papers deal with the case of Bernoulli random variables, but here we are indeed interested in the general discrete case. Few extensions were made in this direction. To the best of our knowledge, Mood (1940) is the first one who studied it.

A run can be defined as follows: it is a consecutive sub-sequence of identical values in a sequence of random numbers. For any $j \in E$, let us denote by $R^i_{j,n}$ the number of non-overlapping j-runs of length i in the sequence $X_1, \ldots, X_n$:

$$R^i_{j,n} = |\{m \ ; \ X_{m-1} \neq j \ , , \ X_m = j \ , , \ \ldots \ , , \ X_{m+i-1} = j \ , , \ X_{m+i} \neq j\}| \ .$$

Let us now define the number $R_{j,n}$ of j-runs and the total number R_n of non-overlapping runs:

$$R_{j,n} = \sum_{i=1}^n R^i_{j,n} \ , \qquad \text{and} \qquad R_n = \sum_{j \in E} R_{j,n} \ .$$

Mood (1940) obtained the limiting distribution of R_n after renormalization. Two cases have be distinguished: $k = 2$ and $k > 2$.

Theorem 4.2 *(corollary 5 p. 390 and corollary 3 p. 392 in (Mood, 1940))*

1. *If* $k = 2$, $\dfrac{R_n - 2n\pi_1\pi_2}{2\sqrt{n\pi_1\pi_2(1-3\pi_1\pi_2)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$.
2. *If* $k > 2$, $\frac{1}{\sqrt{n}}\left(R_n - n\left(1 - \sum_{j\in E}\pi_j^2\right)\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\sigma_\pi^2)$, *with* $\sigma_\pi^2 = \sum_{j\in E}\pi_j^2 + 2\sum_{j\in E}\pi_j^3 - 3(\sum_{j\in E}\pi_j^2)^2$.

One can check that in both cases the asymptotic variance is degenerated if and only if there exists $j \in E$ such that $\pi_j = 1$ (and then, for all $j' \neq j$, $\pi_{j'} = 0$), then the variance is degenerated. These convergence results could be used to construct an asymptotic test.

An alternative solution could be to consider the longest run in the sequence $X_1, \ldots, X_n$. Indeed there exists many works dealing with the case of either independent trials or Markovian trials. Let us denote by L_n the longest length of all runs in the sequence $X_1, \ldots, X_n$. Using the previous notations, we have:

$$L_n = \max\{i \ ; \ \exists j \ s.t. \ R_{j,n}^i > 0\} \ .$$

Vaggelatou (2003) studied this random variable in the case of multi-state Markovian trials. It requires that $\{X_t\}$ is an irreducible and aperiodic Markov chain on a finite state space E with transition probability matrix P and unique stationary measure π. The Markov chain induced by a DAR(1) model is irreducible and aperiodic if $\pi > 0$ (meaning that all the components of π are strictly positive) and $\alpha \neq 1$ (see chapter XV of (Feller, 1968)). Let us define the two following quantities:

$$\rho = \max_{j\in E} P_{jj} \quad \text{and} \quad \pi_\rho = \sum_{j\in E \,:\, P_{jj}=\rho} \pi_j \ .$$

If $\rho < 1$, then Vaggelatou proved the following asymptotic result (theorem 1 in (Vaggelatou, 2003)):

$$\mathbb{P}(L_n - [\log_{1/\rho} n] < x) = \exp\left\{-n(1-\rho)\pi_\rho \rho^{[\log_{1/\rho} n]+x-1}\right\} + o(1) \ , \tag{6}$$

where $[\cdot]$ denotes the integer part and $o(1)$ means that the residual term is ¡¡ small ¿¿ in regard with n. This result extends the classical one obtained many years ago by Gončarov (1962) in the case of iid Bernoulli trials. In both case, $L_n - [\log_{1/\rho} n]$ does not have a limit distribution, but only certain sub-sequence; for instance, theorem 2 in (Vaggelatou, 2003) gives a case where the sub-sequence converges in distribution to the Gumbel distribution. We will use theorem 1 (and not theorem 2) to construct a third (and last) test since we have not enough observations in real situation. Let us denote by ρ_0 and ρ_1 the value of ρ respectively under the null and the alternative hypothesis. Under the null hypothesis, P will be equal to the matrix Q as defined in the introduction: $\forall (j,j') \in E^2$, $P_{jj'} = \pi_{j'}$ (the transition probability from state j to state j' does not depend on j). So we have that $\rho_0 = \max\{\pi_j \ ; \ j \in E\}$: $\rho_0 < 1$ if and only if, for any $j \in E$, $\pi_j < 1$. Under the alternative hypothesis, $\rho_1 = \max\{P_{jj} \ ; \ j \in E\} = \max\{\alpha + (1-\alpha)\pi_j \ ; \ j \in E\}$: $\rho_1 < 1$ if and only if $\alpha < 1$ (it is initially assumed) and for any $j \in E$, $\pi_j < 1$. Thus we find the same condition in both cases and this assumption is the same as for the previous test. From now we will assume that $\pi > 0$ in addition to the previous assumptions (let us recall that we already assume that $\alpha \neq 1$). Using theorem 1 of Vaggelatou (2003), one could obtain an asymptotic confidence interval with a prescribed confidence level $\varepsilon \in (0;1)$:

$$\mathbb{P}_{H_0}\left(\tilde{L}_n \in \bar{W}_{\varepsilon,n} = \left[\log_{\rho_0}\left(-\frac{\ln(\varepsilon/2)}{n(1-\rho_0)\pi_{\rho_0}}\right) \ ; \ \log_{\rho_0}\left(-\frac{\ln(1-\varepsilon/2)}{n(1-\rho_0)\pi_{\rho_0}}\right)\right]\right) = 1-\varepsilon \ ,$$

$\tilde{L}_n = L_n - 1$ (notice that it is corresponding sometimes to the definition of runs: see for instance (Jacobs and Lewis, 1978a)). It follows that the power Π_ε of this test is:

$$\Pi_\varepsilon = 1 + \mathbb{P}_{H_1}\left(\tilde{L}_n < \log_{\rho_0}\left(-\frac{\ln(\varepsilon/2)}{n(1-\rho_0)\pi_{\rho_0}}\right)\right) - \mathbb{P}_{H_1}\left(\tilde{L}_n < \log_{\rho_0}\left(-\frac{\ln(1-\varepsilon/2)}{n(1-\rho_0)\pi_{\rho_0}}\right)\right) \ .$$

To compute Π_ε, one has to use equation (6) above.

4.3 Parameters estimations

The two parameters π and α of such DAR(1) model could be estimated separately since by construction they play different role.

Estimations of π For any $j \in E$ and for any $i \in \{1, \ldots, n\}$, let $Z_{ij} = \mathbb{1}_{\{X_i = j\}}$: these random variables have the Bernoulli distribution with parameter π_j. A natural unbiased estimator of π_j is therefore:

$$\widehat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} Z_{ij} \ .$$

Moreover, using the expression of P^h given in the introduction, one can easily derive the variance of $\widehat{\pi}_j$:

$$\text{Var}[\widehat{\pi}_j] = \frac{1}{n}\pi_j(1-\pi_j) + \frac{2}{n^2}(1-\pi_j)\pi_j V_n(\alpha) \ ,$$

and the covariance between $\widehat{\pi}_j$ and $\widehat{\pi}_{j'}$ (with $j' \neq j$):

$$\text{Cov}[\widehat{\pi}_j, \widehat{\pi}_{j'}] = -\frac{2}{n^2}\pi_{j'}\pi_j V_n(\alpha) \ ,$$

with $V_n(\alpha) = \sum_{h=1}^{n}(n-h)\alpha^h$. Applying formula (0.113) in (Gradshteyn and Ryznik, 1965) (arithmetico-geometric progression), we obtain the following expression for $V_n(\alpha)$:

$$V_n(\alpha) = \frac{n - \alpha^n}{1-\alpha} - \frac{\alpha(1-\alpha^{n-1})}{(1-\alpha)^2} - n \ .$$

This leads to the following limit for the variances and the covariances:

$$\lim_{n\to\infty} n\text{Var}[\widehat{\pi}_j] = \frac{1+\alpha}{1-\alpha}\pi_j(1-\pi_j) \ ,$$

and:

$$\lim_{n\to\infty} n\text{Cov}[\widehat{\pi}_j, \widehat{\pi}_{j'}] = -\frac{2\alpha}{1-\alpha}\pi_{j'}\pi_j \ .$$

As a consequence of Bienaymé-Chebychev inequality, the asymptotic result on the variance implies that $\widehat{\pi}_j$ is consistent:

Proposition 4.1 *For any $\alpha \in [0,1)$, $\widehat{\pi}_j \xrightarrow[n\to\infty]{Pr} \pi_j$.*

Moreover one can prove the following central limit theorem for $\widehat{\pi}_j$ by application of the ergodic theorem for Markov chain (Jones, 2004) and Slutsky theorem:

Theorem 4.3 *For any $\alpha \in [0,1)$, $\sqrt{n}\dfrac{\widehat{\pi}_j - \pi_j}{\sqrt{\widehat{\pi}_j(1-\widehat{\pi}_j)}} \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \dfrac{1+\alpha}{1-\alpha}\right)$.*

When $\alpha = 0$, the asymptotic variance equals to 1 as it is well-known for Bernoulli trials. The largest α is, the larger the asymptotic variance is. Since the asymptotic depends on α which is generally unknown, we cannot yet use the last proposition to construct confidence interval. In order to do it, we will need an consistent estimator of α.

Estimations of α We will first consider the maximum likelihood estimator of α, assuming that π is known. Since $\{X_t\}$ is a Markov chain, the log-likelihood is:

$$\mathcal{L}(X_1, \ldots, X_n; \alpha) = \sum_{(j,j') \in E^2} N_{j,j'} \log P_{j,j'}(\alpha') \ ,$$

where $N_{j,j'}$ is defined as in section 2 and where $P_{j,j'}(\alpha)$ is the transition probability from state j to state j' (which only depends on α). Replacing the expression of transition probabilities, we obtain that:

$$\mathcal{L}(X_1,\ldots,X_n;\alpha) = \sum_{j\in E}\left(N_{j,j}\log(\alpha+(1-\alpha)\pi_j) + \sum_{j'\in E\setminus\{j\}} N_{j,j'}\log((1-\alpha)\pi_{j'})\right) .$$

It follows that the maximum likelihood estimator α_1^* of α is the solution of the following equation:

$$\frac{1}{n}\sum_{j\in E}\frac{N_{j,j}^n}{\alpha+(1-\alpha)\pi_j} = 1 .$$

When π is unknown, one can use the estimation given above and so the plug-in estimator $\widehat{\alpha_1}$ of α is the solution of the following equation:

$$\frac{1}{n}\sum_{j\in E}\frac{N_{j,j}^n}{\alpha+(1-\alpha)\widehat{\pi}_j} = 1 .$$

Unfortunately we cannot derive an explicit expression of $\widehat{\alpha_1}$. An alternate possible way is to minimize the following function:

$$Q_\alpha = \sum_{(j,j')\in E^2}(\widehat{P}_{j,j'} - P_{j,j'}(\alpha))^2 .$$

Solving this optimization problem leads to an explicit expression ${\alpha_2}^*$:

$$\alpha_2^* = \frac{\sum_{j\in E}(1-\pi_j)(\widehat{P}_{jj}-\pi_j) - \sum_{j\in E}\sum_{j'\in E\setminus\{j\}}\pi_{j'}(\widehat{P}_{jj'}-\pi_{j'})}{(k-1)\sum_{j\in E}\pi_j^2 + \sum_{j\in E}(1-\pi_j)^2} .$$

It seems to be difficult to establish properties of this intuitive estimator. Hence it is not recommended to use it as an estimator of α. However it could provide a possible initialization for an optimization procedure to obtain a numerical value of $\widehat{\alpha_1}$. The corresponding plug-in estimator $\widehat{\alpha_1}$ of α is given by the following expression:

$$\widehat{\alpha_2} = \frac{\sum_{j\in E}(1-\widehat{\pi}_j)(\widehat{P}_{jj}-\widehat{\pi}_j) - \sum_{j\in E}\sum_{j'\in E\setminus\{j\}}\widehat{\pi}_{j'}(\widehat{P}_{jj'}-\widehat{\pi}_{j'})}{(k-1)\sum_{j\in E}\widehat{\pi}_j^2 + \sum_{j\in E}(1-\widehat{\pi}_j)^2} .$$

Simulated data The estimators developed above are applied on simulated data in order to evaluate numerically their performance. We simulate data with various values of α, π and n:

- $\pi = (\frac{1}{2},\frac{1}{2})$, $\pi = (\frac{1}{3},\frac{2}{3})$, $\pi = (\frac{1}{3},\frac{1}{3},\frac{1}{3})$ and $\pi = (\frac{1}{4},\frac{1}{2},\frac{1}{4})$
- $\alpha \in \{0.1; 0.2; 0.5; 0.8; 0.9\}$.
- $n \in \{50; 100; 500\}$.

Hence two cases are studied: $k = 2$ and $k = 3$. For each values of these parameters, we simulate $m = 100$ independent DAR(1) Markov chain and then we compute the estimators. Unfortunately these we have no guarantee that the two estimators of α belong to the unit interval. Hence we precise for each cases the number of samples on which the computations were done (as it is reasonable, the number of samples is increasing with the number n of observations). Results are given in tables 1 to 4 (only the estimators are computed for simulated data).

4.4 A variant with missing observations

Sometimes categorical time series may contain some missing values/observations. Here we now propose a very simple adaptation of the DAR model in order to taking into account the missing values. Since the DAR model is

stationary, it will be easy to derive similar expressions as in the initial model.

Assume that at each unit of time, the probability of a missing value equals to β (which does not depend on the time). Hence, if we denote by Z_t the values at time t, we have :

$$Z_t = \begin{cases} X_t & \text{w.p.} \quad 1-\beta \\ -1 & \text{w.p.} \quad \beta \end{cases} ,$$

where -1 is the value corresponding to a missing value and $\{X_t\}$ is DAR(1) stochastic process as described previously. Hence β is the probability that a value is not observed : this probability is assumed to be not depending on t. Since (X_t) is a stationary stochastic process, it follows that (Z_t) is still a Markov chain, but taking values on the set $\tilde{E} = \{-1\} \cup E$. Its transition probabilities matrix $\tilde{P}$ can be expressed in function of the transition probabilities matrix P of (X_t):

$$\tilde{P} = \begin{bmatrix} 1-\beta & \beta^t \pi \\ (1-\beta)\mathbf{1}_k & \beta P \end{bmatrix} ,$$

where $\mathbf{1}_k$ is the unit vector of $\mathbb{R}^k$. There is now three parameters to be estimated. Indeed π and α (with the maximum likelihood method) can be estimated as previously. The extra parameter β can be simply estimated as follows:

$$\widehat{\beta} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i = -1\}} .$$

5 Applications to ecological data

Finally tests and estimators are applied in this section to real data. We apply the two models described previously to two real data sets. The first one deals with larch cone production (see section 2) and the second one to planktonic abundance.

5.1 Larch cone production

The total number of observations was $n = 31$, with $k = 6$ categories. The goal of this work is to study masting on such trees. For a full description of the data, see section 2.

First we apply regression models. Tables 5 and 6 contain values of AIC respectively for the categorical time series regression model and the ordinal one. We also indicate the number of parameters to estimate and the number of observations (these time series may contain missing values). For the first case, the independence assumption leads to the better for all sites. For the second case, model with a lag order 1 and 2 fits better for the site Ayes 2200 and Montgenèvre 2200 while the model with only a lag of order 1 fits better for Névache 1800 and the independent model fits better for Prorel 1800. Comparing values of AIC, the model for ordinal time series seems to be more accurate for these data sets.

Second we apply the DAR model. Table 7 contains the estimations of the three parameters for each of the four data sets. In all cases, the independence hypothesis is rejected with the two first tests, while the third one leads to accept the assumption of independence (all with a first type error at 5%). However the power of this last test is more or less weak in all cases (ranging from above 26% to 51%). Thus it is reasonable to reject the assumption of independent observations.

In all cases the categorical time series regression model leads to accept the temporal independence between observations. It may be due to the fact that the parameter α in the DAR model is closed to zero. However this parameter can be assumed to be significantly different of zero, according to the performed tests. It is in concordance with the fact that observations are time-dependent when using the ordinal time series regression model. Time series studied here are very short-length and it may the cause that the conclusions based on the ordinal time series regression model and the ones based on the DAR models. Since few data are available, one should prefer to use the DAR models (because it involves less parameters than the other models).

5.2 Planktonic abundance

We now consider weekly planktonic (*Thalia democratica*) abundance data. Data were kindly given by F. Ménard. In such context, the objective is to test and to compare the temporal patterns from one year to another. Hence we apply the two models described above for four years (1987 to 1990). It follows that each data-set is made of $n = 52$ observations. Abundances were determined semi-quantitatively according to classes defined on scale of 5 values, $E = \{1, 2, 3, 4, 5\}$. The observed series are shown in figure 2 and exhibit the same problems as the larch cone production ones. Notice that the fifth category were not observed for any year. For a complete description of the data, the reader could reefer to (Ménard *et al.*, 1993) (see also (Viennet *et al.*, 1998)).

First we apply regression models. Tables 8 and 9 contain the number of parameters to be estimated, the values of AIC and the number of observations, respectively for the categorical time series regression model and the ordinal one. With the model for categorical time series, the model with one order lag fits better all the four years. With the model for categorical time series, model with two order lag fits better for the year 1987 and 1989 while model with only a one order lag fits better for the two other years, 1988 and 1990.

Second we apply the DAR model. Table 10 contains the estimations of the three parameters for each of the four data sets. The two first tests leads to reject the null hypothesis, i.e. to reject the independence of the observations. The third test leads also to reject the independence assumption for the two last year (1989 and 1990) while the null hypothesis is accepted according to this last test for the years 1987 and 1988 (respectively with a power equal to 44% and 66%). Thus it is also reasonabke to reject the assumption of independent observations.

We can also analyze these data sets as one unique time series (notice that it was not possible for the previous data-set). Hence the sample size is now $n = 208$. All the tests lead to reject the assumption of independent observations (with a power equal to 69% for the last one). The last line of table 10 contains the estimations of the three parameters for the whole period.

Here the situation is totally different than previously. Indeed the categorical time series regression model and the DAR model lead to the same conclusion, i.e. a one order lag dependence in the time series. One can notice that for these data sets the parameter α (of the DAR model) is now between 0.308 and 0.540. However the ordinal time series regression model leads in some case to a two order lag dependence. Since the number of observations is almost the twice than for the first data sets, one should rather prefer to use the ordinal time series regression model.

6 Conclusions and discussions

Applications to real data achieve to convince that these two complementary models are relevant for practical purpose. Based on the result obtained over real data, one can conclude that either the ordinal time-series regression model or the DAR model should be used to treat such data. Indeed in all cases the categorical time-series regression model seems not to present advantages over the two other models. The choice between the ordinal time-series regression model and the DAR model depends highly on the context, i.e. essentially on the number of observations and the number of parameters to be estimated. Fokianos and Kedem (2003) claimed that ¡¡ the regression methodology can discover dependencies in the DNA sequence data which cannot assessed by a Markov model ¿¿. However data treated here can serve as a counter-example of this sentence. For the two data sets studied here, conclusions based on the regression models and the one based on the Markov model are almost identical.

From this work, one can conclude that in any case the DAR model has only few parameters to be estimated, but with an equal to number of unknown parameters (as in the example of larch cone production) one has to prefer the ordinal time-series regression model.

However both suffers of relying on assumptions or simplifications. Hence these models could be extended in the following ways:

- *Stationarity:* the DAR model is strongly stationary (in the sense defined by McGee and Harris in (2005)). This assumption should be checked with any statistical tests. However no test of stationarity of a categorical time series has been developed to the best of our knowledge. Anyway when dealing with short-length time series stationary assumption is not really restrictive. Otherwise a solution could be in applying the de-trend algorithm suggested by McGee and Harris in (2005). However their algorithm is more and more computationally complex as the number of states is increasing (in fact they mainly consider the binary case). We do not focus here on the study of the possible stationarity of a categorical time series which will be done in a future work. A major advantage of regression model is that it is not necessary to have stationarity.
- *Higher dependence order and number of parameters:* since we consider the case of short-length data, we limit our study to one or two order lagged models. Indeed both models could be applied to p-th order lagged models. However, for the regression model, a large value of p implies a large number of parameters to estimate, that may induce some numerical instability (due to correlation between the regressors). The number of parameters in a DAR(p) model is lower, but if $p > 1$ we have no more the Markov property.
- *Environmental factors:* these models do not include environmental covariates. The regression model could easily integrate such situations, as shown by Fokianos and Kedem (2002; 2003). For the DAR model, it is not so easy. A solution could be to consider inhomogeneous Markov chain or a state-space model.

Acknowledgments We wish to thank Patricia Jacobs (Naval Postgraduate School, Monterey, California, USA), Ian R. Harris (Southern Methodist University, Texas, USA), Alain Latour (LabSAD, Grenoble, France), Monnie McGee (Southern Methodist University, Texas, USA) and Frédéric Ménard (IRD, Montpellier, France). Thanks also to the numerous scientists and crew members who conducted the experimental campaigns because our analysis is based on their hard work. This work is a contribution to the understanding of larch cone production into an IFB (Institut Français de la Biodiversité) project.

References

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley, New York.

Bandt, C. (2005). Ordinal time series analysis. *Ecological Modelling*, **182**, 229–235.

Chalmers, J. and Hastie, T. (1992). *Statistical Models in S*. Wadsworth & Brooks, Cole Advanced Books & Software.

Feller, W. (1968). *An introduction to probability theory and its applications: volume I*. John Wiley, New York.

Fokianos, K. and Kedem, B. (2002). *Regression model for time series analysis*. Wiley Interscience.

Fokianos, K. and Kedem, B. (2003). Regression theory for categorical time series. *Statistical science*, **18**(3), 357–376.

Gončarov, V. (1962). On the field of combinatory analysis. *American Mathematical Society Translations*, **19**(2), 1–46.

Gradshteyn, I. and Ryznik, I. (1965). *Academic Press, New-York*. Table on integrals series and products.

Jacobs, P. and Lewis, P. (1978a). Discrete time series generated by mixtures i: correlational and runs properties. *Journal of the Royal Statistical Society (Series B)*, **40**(1), 94–105.

Jacobs, P. and Lewis, P. (1978b). Discrete time series generated by mixtures ii: Asymptotic properties. *Journal of the Royal Statistical Society (Series B)*, **40**(2), 222–228.

Jacobs, P. and Lewis, P. (1978c). Discrete time series generated by mixtures iii: autoregressive processes. Technical Report NPS55-78-022, Naval Postgraduate School, Monterey, CA.

Jacobs, P. and Lewis, P. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis*, **4**(1), 19–36.

Jones, G. (2004). On the markov chain central limit theorem. *Probability Surveys*, **1**, 299–320.

Kauffmann, H. (1987). Regression models for nonstationary categorical time series: asymptotic estimation theory. *Annals of Statistics*, **15**(1), 79–98.

Kelly, D. and Sork, V. (2002). Mast seeding in perennial plants: why, how, where? *Annual Review of Ecology and Evolution and Systematics*, **33**, 427–447.

Liebhold, A., Koenig, W. D., and Bjørnstad, O. N. (2004). Spatial synchrony in population dynamics. *Annual Review of Ecology and Evolution and Systematics*, **35**, 467–490.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society (Series B)*, **42**, 109–142.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman and Hall, London.

McGee, M. and Harris, I. (2005). Coping with nonstationarity in categorical time series. Technical Report 319, Southern Methodist University.

McKenzie, E. (2003). Discrete variate time series. In D. S. et al., editor, *Stochastic processes: modelling and simulation*, volume 21 of *Handbook of Statistics*, pages 573–606. North-Holland.

Ménard, F., Dallot, S., and Thomas, G. (1993). A stochastic model for ordered categorical time series. application to planktonic abundance data. *Ecological Modelling*, **66**, 101–112.

Mood, A. (1940). The distribution theory of runs. *Annals of Mathematical Statistic*, **11**, 367–392.

Price, B., Allgöwer, B., and Fischlin, A. (2006). Synchrony and travelling waves of larch bud moth? time series analysis with changing scale. *Ecological Modelling*, **199**, 433–441.

Raftery, A. (1985). A model for high-order markov chains. *Journal of the Royal Statistical Society (Series B)*, **47**, 528–539.

Reinert, G., Schbath, S., and Waterman, M. (2000). Probabilistic and statistical properties of words: an overview. *Journal of Computational Biology*, **7**(1-2), 1–46.

Roques, A. (1988). The larch cone fly in the french alps. In A. Berryman, editor, *Dynamics of forest insect populations: patterns, causes, implications*. Plenum, Washington.

Vaggelatou, E. (2003). On the length of the longest run in a multi-state markov chain. *Statistic and Probability Letters*, **62**(3), 211–221.

Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S*. Springer.

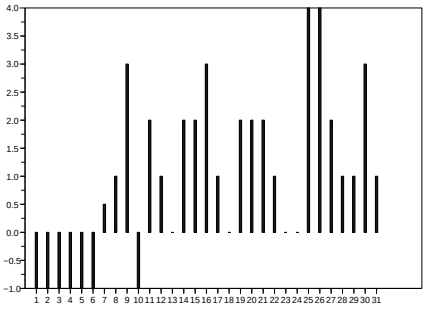
Viennet, G., Ménard, F., and Thomas, G. (1998). Partial likelihood estimation in categorical time series with stochastic covariates. *Biometrics*, **54**, 304–311.

Biographical sketches

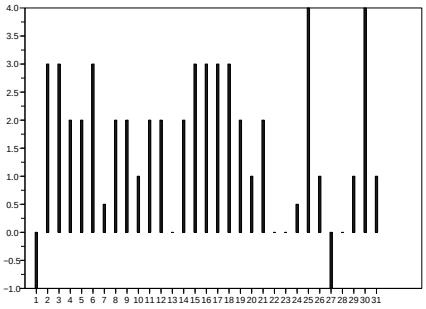
Noëlle Bru is an Assistant Professor at the IUT STID of Université de Pau et des Pays de l'Adour and a Researcher at the Laboratoire de Mathématiques Appliqués de Pau. Since her PhD thesis, her topics of interest is both theoretical and applied statistics with emphasis to environmental data analysis.

Laurence Despres is an Assistant Professor at the Université Joseph Fourier and a Researcher at the Laboratoire d'Ecologie Alpine. Her main research interests are in the evolutionary ecology of species interactions and coevolution.

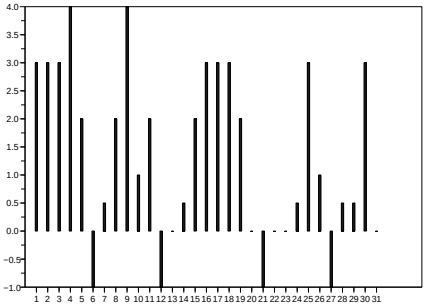
Christian Paroissin is an Assistant Professor at the Département Sciences et Techniques of Université de Pau et des Pays de l'Adour and a Researcher at the Laboratoire de Mathématiques Appliqués de Pau. His topics of interest is applied probability and statistics with interest to applied contexts (engineering, theoretical computer science, biology, ...).



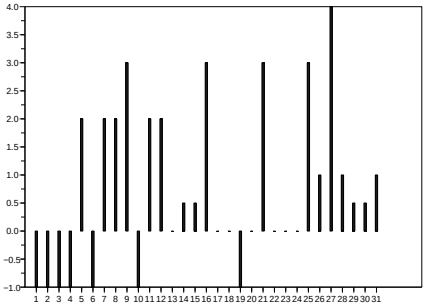
(a) Ayes 2200



(b) Montgenèvre 2200



(c) Névache 1800

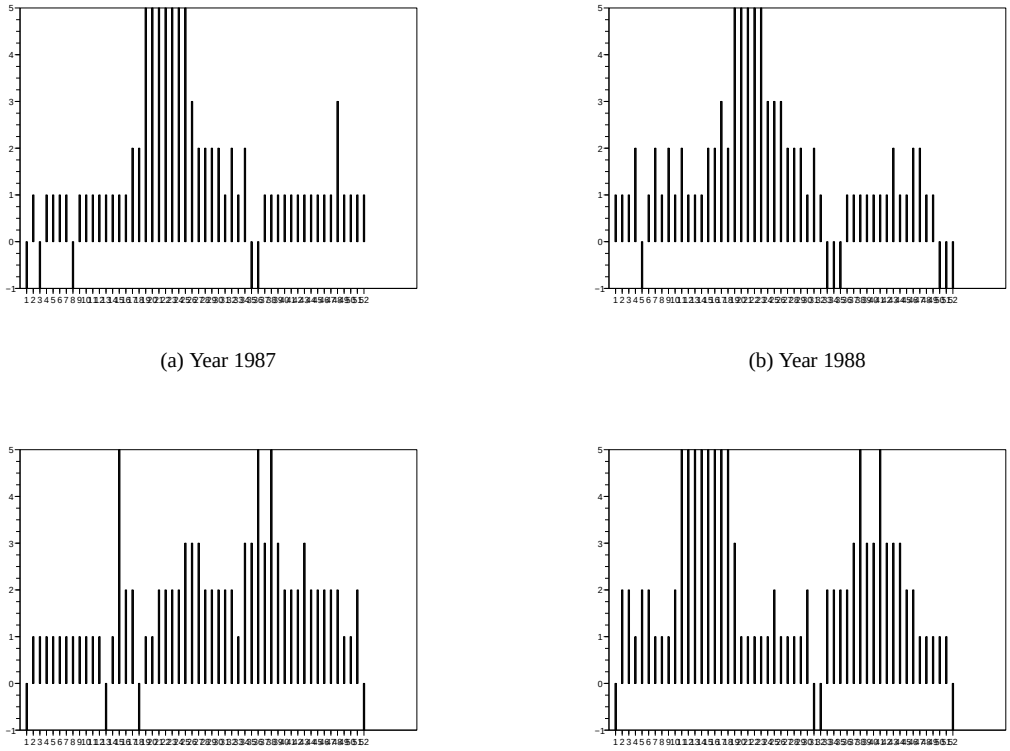


(d) Prorel 1800

Figure 1: Annual larch production in four sites in the Southern Alps

α	n	$\widehat{\pi}$	$\widehat{\alpha_1}$	m_1	$\widehat{\alpha_2}$	m_2
0.1	50	(0.509;0.491)	0.140	65	0.149	72
	100	(0.499;0.501)	0.114	82	0.121	85
	500	(0.503;0.497)	0.092	100	0.094	100
0.2	50	(0.511;0.489)	0.174	83	0.183	90
	100	(0.510;0.490)	0.180	97	0.192	97
	500	(0.499;0.501)	0.191	100	0.194	100
0.5	50	(0.501;0.499)	0.451	99	0.481	99
	100	(0.510;0.490)	0.465	100	0.479	100
	500	(0.499;0.501)	0.494	100	0.497	100
0.8	50	(0.543;0.457)	0.711	99	0.753	99
	100	(0.534;0.466)	0.765	100	0.784	100
	500	(0.502;0.498)	0.795	100	0.799	100
0.9	50	(0.585;0.415)	0.754	99	0.798	99
	100	(0.539;0.461)	0.857	100	0.882	100
	500	(0.503;0.494)	0.890	100	0.894	100

Table 1: Results obtained with $\pi = (\frac{1}{2}, \frac{1}{2})$



(a) Year 1987

(b) Year 1988

(c) Year 1989

(d) Year 1990

Figure 2: Weekly planktonic abundance for four years

α	n	$\widehat{\pi}$	$\widehat{\alpha_1}$	m_1	$\widehat{\alpha_2}$	m_2
0.1	50	(0.328;0.672)	0.145	65	0.160	69
	100	(0.337;0.663)	0.120	74	0.131	15
	500	(0.337;0.663)	0.095	97	0.097	97
0.2	50	(0.335;0.665)	0.202	91	0.226	91
	100	(0.335;0.665)	0.205	96	0.217	96
	500	(0.336;0.664)	0.190	100	0.192	100
0.5	50	(0.357;0.643)	0.442	100	0.471	100
	100	(0.332;0.668)	0.461	100	0.475	100
	500	(0.334;0.666)	0.486	100	0.489	100
0.8	50	(0.364;0.636)	0.711	99	0.744	99
	100	(0.352;0.648)	0.746	99	0.764	99
	500	(0.343;0.657)	0.790	100	0.793	100
0.9	50	(0.438;0.562)	0.795	93	0.846	94
	100	(0.410;0.590)	0.849	100	0.870	100
	500	(0.336;0664)	0.893	100	0.896	100

Table 2: Results obtained with $\pi = (\frac{1}{3}, \frac{2}{3})$

α	n	$\widehat{\pi}$	$\widehat{\alpha_1}$	m_1	$\widehat{\alpha_2}$	m_2
0.1	50	(0.324;0.345;0.331)	0.120	74	0.132	75
	100	(0.337;0.339;0.324)	0.099	82	0.103	84
	500	(0.333;0.329;0.338)	0.104	100	0.105	100
0.2	50	(0.345;0.334;0.321)	0.176	95	0.184	98
	100	(0.327;0.342;0.331)	0.186	99	0.193	99
	500	(0.330;0.336;0.334)	0.199	100	0.201	100
0.5	50	(0.328;0.315;0.357)	0.443	100	0.447	100
	100	(0.345;0.336;0.319)	0.477	100	0.483	100
	500	(0.340;0.323;0.332)	0.495	100	0.496	100
0.8	50	(0.366;0.309;0.325)	0.737	100	0.713	100
	100	(0.382;0.309;0.309)	0.756	100	0.758	100
	500	(0.336;0.328;0.336)	0.793	100	0.795	100
0.9	50	(0.463;0.254;0.283)	0.776	99	0.722	100
	100	(0.411;0.283;0.306)	0.860	100	0.837	100
	500	(0.348;0.322;0.330)	0.895	100	0.893	100

Table 3: Results obtained with $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

α	n	$\widehat{\pi}$	$\widehat{\alpha_1}$	m_1	$\widehat{\alpha_2}$	m_2
0.1	50	(0.255;0.492;0.253)	0.114	82	0.128	82
	100	(0.248;0.495;0.257)	0.097	89	0.102	90
	500	(0.249;0.501;0.250)	0.098	100	0.099	100
0.2	50	(0.265;0.484;0.251)	0.174	92	0.179	94
	100	(0.255;0.494;0.251)	0.177	99	0.182	99
	500	(0.246;0.506;0.248)	0.194	100	0.195	100
0.5	50	(0.279;0.467;0.254)	0.467	100	0.467	100
	100	(0.250;0.494;0.256)	0.466	100	0.467	100
	500	(0.253;0.497;0.250)	0.493	100	0.493	100
0.8	50	(0.299;0.496;0.205)	0.716	100	0.693	100
	100	(0.282;0.472;0.246)	0.766	100	0.765	100
	500	(0.258;0.503;0.239)	0.799	100	0.799	100
0.9	50	(0.356;0.422;0.222)	0.802	100	0.715	100
	100	(0.332;0.460;0.208)	0.858	100	0.827	100
	500	(0.248;0.493;0.259)	0.893	100	0.892	100

Table 4: Results obtained with $\pi = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$

		Ayes 2200		Montgenèvre 2200		Névache 1800		Prorel 1800	
Model	Nb param	AIC	Nb obs	AIC	Nb obs	AIC	Nb obs	AIC	Nb obs
Indep.	6	**119.63**	24	**143.25**	29	**143.36**	27	**128.43**	24
Lag 1	42	141.26	22	157.58	27	169.06	22	154.02	20
Lags 1-2	78	180.27	20	187.00	25	212.72	17	179.23	17

Table 5: Categorical time-series regression models applied to annual larch cones production

		Ayes 2200		Montgenèvre 2200		Névache 1800		Prorel 1800	
Model	Nb param	AIC	Nb obs	AIC	Nb obs	AIC	Nb obs	AIC	Nb obs
Indep.	6	119.63	24	143.25	29	143.36	27	**128.43**	24
Lag 1	12	116.34	22	136.3	27	**139.32**	22	136.62	20
Lags 1-2	18	**114.3**	20	**126.92**	25	150.13	17	144.82	17

Table 6: Ordinal time-series regression models applied to annual larch cones production

Valley	$\widehat{\pi}$	$\widehat{\alpha_1}$	$\widehat{\beta}$	AIC
Ayes 2200	(0.167;0.042;0.292;0.292;0.125;0.083)	0.082	0.774	**121.06**
Montgenèvre 2200	(0.138;0.069;0.172;0.310;0.241;0.069)	0.070	0.935	**118.70**
Névache 1800	(0.185;0.185;0.074;0.185;0.296;0.074)	0.032	0.871	**125.94**
Prorel 1800	(0.292;0.167;0.125;0.208;0.167;0.042)	0.161	0.774	**122.89**

Table 7: DAR models applied to annual larch cones production

		1987		1988		1989		1990		1987-1990	
Model	Nb param	AIC	Nb obs	AIC	Nb obs	AIC	Nb obs	AIC	Nb obs	AIC	Nb obs
Indep.	3	100.56	47	111.52	45	121.61	48	133.6	48	468.21	188
Lag 1	12	**73.04**	43	**89.83**	42	**96.06**	45	**99.69**	46	331.81	177
Lags 1-2	21	79.52	40	93.57	37	97.17	42	102.90	44	**315.96**	167

Table 8: Categorical time-series regression models applied to weekly planktonic abundance

		1987		1988		1989		1990		1987-1990	
Model	Nb param	AIC	Nb obs	AIC	Nb obs	AIC	Nb obs	AIC	Nb obs	AIC	Nb obs
Indep.	3	100.56	47	111.52	45	121.61	48	133.6	48	468.21	188
Lag 1	6	65.77	43	**84.78**	42	95.67	45	**92.66**	46	323.5	177
Lags 1-2	9	**63.69**	40	NA	37	**85.92**	42	93.65	44	**298.17**	167

Table 9: Ordinal time-series regression models applied to weekly planktonic abundance

Year	$\widehat{\pi}$	$\widehat{\alpha_1}$	$\widehat{\beta}$	AIC
1987	(0.625;0.167;0.042;0.000;0.167)	0.540	0.923	**108.78**
1988	(0.489;0.311;0.089;0.000;0.111)	0.308	0.865	**143.11**
1989	(0.354;0.417;0.167;0.000;0.062)	0.445	0.923	**132.68**
1990	(0.375;0.271;0.146;0.000;0.208)	0.484	0.923	**135.10**
1987-1990	(0.463;0.293;0.112;0.000;0.133)	0.468	0.904	**511.00**

Table 10: DAR models applied to weekly planktonic abundance